ICS 11.040.99
CCS C 30

YY

中华人民共和国医药行业标准

YY/T 1833.1—2022

人工智能医疗器械　质量要求和评价 第1部分：术语

Artificial intelligence medical device—Quality requirements and evaluation—Part 1：Terminology

2022-07-01 发布　　2023-07-01 实施

国家药品监督管理局　　发布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

本文件是YY/T 1833《人工智能医疗器械　质量要求和评价》的第1部分。YY/T 1833已经发布了以下部分：

——第1部分：术语；

——第2部分：数据集通用要求。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由国家药品监督管理局提出。

本文件由全国人工智能医疗器械标准化技术归口单位归口。

本文件起草单位：中国食品药品检定研究院、解放军总医院、中国科学院深圳先进技术研究院、上海长征医院、上海联影智能医疗科技有限公司、飞利浦（中国）投资有限公司、上海西门子医疗器械有限公司、杭州依图医疗技术有限公司、腾讯医疗健康（深圳）有限公司、华为技术有限公司、广州柏视医疗科技有限公司、美中互利医疗有限公司、慧影医疗科技（北京）有限公司、北京安德医智科技有限公司、中山大学中山眼科中心、东南大学。

本文件主要起草人：孟祥峰、何昆仑、郑海荣、刘士远、李佳戈、王浩、詹翊强、王晨希、萧毅、葛鑫、刘东泉、颜子夜、钱天翼、符海芳、陆遥、刘毅、柴象飞、王珊珊、周娟、张培芳、林浩添、万遂人。

引　言

近年来，人工智能医疗器械不断发展，成为医疗器械标准化领域的一个新兴方向。我国已初步建立人工智能医疗器械标准体系。在该标准体系中，YY/T 1833《人工智能医疗器械　质量要求和评价》是基础通用标准，为开展细分领域的标准化活动提供指导，拟由八个部分组成。

——第1部分：术语。目的在于为人工智能医疗器械的质量评价活动提供术语。

——第2部分：数据集通用要求。目的在于提出数据集的通用质量要求与评价方法。

——第3部分：数据标注通用要求。目的在于提出数据标注环节的质量要求与评价方法。

——第4部分：可追溯性通用要求。目的在于提出人工智能医疗器械可追溯性的通用要求及评价方法。

——第5部分：算法安全要求。目的在于规范人工智能医疗器械采用的人工智能算法的安全要求与评价方法。

——第6部分：环境要求。目的在于规范人工智能医疗器械的运行环境条件要求与评价方法。

——第7部分：隐私保护要求。目的在于加强人工智能医疗器械保护受试者隐私的能力。

——第8部分：伦理要求。目的在于从技术层面实现人工智能伦理的要求，保护人的权益。

本文件为其他部分提供基础、共性的术语。其他部分在起草过程中，可能根据需要，补充特定语境下的专用术语。

人工智能医疗器械　质量要求和评价
第1部分：术语

1　范围

本文件界定了人工智能医疗器械质量要求和评价使用的术语和定义。

本文件适用于人工智能医疗器械。

2　规范性引用文件

本文件没有规范性引用文件。

3　术语和定义

3.1　基础技术术语

3.1.1

人工智能　artificial intelligence；AI

表现出与人类智能（如推理和学习）相关的各种功能的功能单元的能力。

[来源：GB/T 5271.28—2001，28.01.02]

3.1.2

人工智能医疗器械　artificial intelligence medical device；AIMD

采用AI技术实现其预期用途的医疗器械。

注1：如采用机器学习、模式识别、规则推理等技术实现其医疗用途的独立软件。

注2：如采用内嵌AI算法、AI芯片实现其医疗用途的医疗器械。

3.1.3

医疗器械软件　medical device software

在集成到正在开发的医疗器械中的已开发的软件系统，或者预期作为医疗器械使用的软件系统。

注：医疗器械软件包括软件组件和独立软件，软件组件是指嵌入到医疗器械中或作为医疗器械组成部分的软件，独立软件是指具有一个或者多个医疗目的，无需医疗器械硬件即可完成自身预期目的，运行于通用计算平台的软件。

[来源：YY/T 0664—2020，3.11]

3.1.4

模式识别　pattern recognition

通过功能单元对某一对象物理或抽象的模式以及结构和配置的辨识。

[来源：GB/T 5271.28—2001，28.01.13]

3.1.5

人工神经网络　artificial neural network；ANN

由加权链路且权值可调整连接的基本处理元素的网络，通过把非线性函数作用到其输入值上使每

个单元产生一个值，并把它传送给其他单元或把它表示成输出值。

注：也可称为神经网络。

[来源：GB/T 5271.34—2006，34.01.06]

3.1.6

推理　inference

从已知前提导出结论的方法。

注 1：在人工智能领域中，前提是事实或规则。

注 2：术语“推理”既指过程也指结果。

[来源：GB/T 5271.28—2001，28.03.01，有修改]

3.1.7

特征　feature

能表达模式本质的功能或结构特点的可度量属性。

注：如大小、纹理、形状等。好的特征能使同类模式聚类、不同类模式分离。

3.1.8

机器学习　machine learning

功能单元通过获取新知识或技能，或通过整理已有的知识或技能来改进其性能的过程。

注：也可称为自动学习。

[来源：GB/T 5271.31—2006，31.01.02，有修改]

3.1.9

深度学习　deep learning

通过训练具有多个隐层的神经网络来获得输入输出间映射关系的机器学习方法。

3.1.10

监督学习　supervised learning

一种学习策略，获得的知识的正确性通过来自外部知识源的反馈加以测试的学习策略。

注：也可称为监督式学习。

[来源：GB/T 5271.31—2006，31.03.08，有修改]

3.1.11

无监督学习　unsupervised learning

一种学习策略，它在于观察并分析不同的实体以及确定某些子集能分组到一定的类别里，而无需在获得的知识上通过来自外部知识源的反馈，以实现任何正确性测试。

注 1：一旦形成概念，就对它给出名称，该名称就可以用于其他概念的后续学习了。

注 2：也可称为无师(式)学习。

[来源：GB/T 5271.31—2006，31.03.09，有修改]

3.1.12

强化学习　reinforcement learning

一种学习策略，它强调从环境状态到动作映射的过程，目标是使动作从环境中获得的累积奖赏值最大。

3.1.13

半监督学习　semi-supervised learning

一种学习策略，它自行利用少量的具有标记信息的样本和大量没有标记的样本进行学习的框架。

3.1.14

自监督学习　self-supervised learning

一种学习策略，通过基于数据本身设计和建立的各种标记信息来对数据本身的特征、特性进行学习，进而把学习到的数据特征网络作为主干网络迁移到对目标任务的学习中。

3.1.15

弱监督学习　weakly supervised learning

一种学习策略，通过使用有噪声的、不完全的、不精确的外部信息源进行机器学习。

注：该方法减少了对标注数据质量和数量的要求。

3.1.16

集成学习　ensemble learning

通过结合多个学习器来解决问题的一种机器学习范式。

注：其常见形式是利用一个基学习算法从训练集产生多个基学习器，然后通过投票等机制将基学习器进行结合。

3.1.17

主动学习　active learning

学习过程中由学习器挑选未标记样本，并请求外界提供标记信息，其目标是使用尽可能少的查询来取得好的学习性能。

3.1.18

迁移学习　transfer learning

利用一个学习领域 A 上有关学习问题 T(A)的知识，改进学习领域 B 上相关学习问题 T(B)的学习算法的性能。

3.1.19

联邦学习　federated learning

一种从多个数据源协同建立模型的机器学习框架，本地数据访问受限，各个数据源方独立进行本地数据处理，通过交换数据模型，共同建立其学习模型，并将输出结果反馈给用户。

3.1.20

训练　training

基于机器学习算法，利用训练数据，建立或改进机器学习模型参数的过程。

3.1.21

交叉验证　cross validation

一种利用已知数据集获取学习器最优参数，以期望在未知数据集上获得最佳泛化性能。

注：常见的有留一法和 K 重交叉验证法。

3.1.22

过拟合　overfitting

学习器对训练样本过度学习，导致训练样本中不具有普遍性的模式被学习器当作一般规律，降低了泛化性能；典型表现是训练集上的性能越高，测试集上的性能越低。

3.1.23

欠拟合　underfitting

学习器对训练样本学习不充分，导致训练样本中包含的重要模式没有被学习器获取，降低了泛化性能；典型表现是训练集上的性能可以继续提高，测试集上的性能同时得以提高。

3.1.24

前馈网络　feedforward network

在给定层内的各人工神经元之间既没有反馈路径也没有任何路径的多层网络。

注：也可称为前向传播网络、非循环网络。

[来源：GB/T 5271.34—2006，34.02.25，有修改]

3.1.25

反向传播网络　back-propagation network

一种多层网络，它使用反向传播，以便学习期间的连接权调整。

注：也可称为反馈传播网络。

[来源：GB/T 5271.34—2006，34.02.30，有修改]

3.1.26

前馈传播 feedforward propagation

在多层网络中，从输入层朝向网络的输出逐层进行连接权调整的传播。

注：也可称为前向传播。

[来源：GB/T 5271.34—2006，34.03.16，有修改]

3.1.27

反馈传播　feedback propagation

在多层网络中，从连接权调整输出层朝向网络的输入的逐层传播。

注：也可称为反向传播。

[来源：GB/T 5271.34—2006，34.03.17，有修改]

3.1.28

医学知识库　medical knowledgebase

帮助临床科研人员、医务人员等快速、便捷地获取疾病诊断、治疗、用药等全面、系统、动态的临床医学知识的集合，也可作为临床教学及临床诊疗的辅助参考工具。

3.1.29

算法服务　algorithm service

算法在运行状态下提供的推理服务。

注：算法服务接受用户的应用请求，对输入数据进行处理，返回处理结果。

3.1.30

云服务　cloud service

通过云计算已定义的接口，部署在云端服务器上提供的一种或多种能力。

[来源：GB/T 32400—2015，3.2.8]

3.1.31

边缘云服务　edge cloud service

通过云计算已定义的接口，部署在边缘计算设施上提供的一种或多种能力。

3.1.32

本地服务　local service

部署在终端或本地计算设施上提供的一种或多种能力。

3.1.33

人工智能医疗器械生存周期模型　AIMD lifecycle model

人工智能医疗器械从起始到停运的整个演进过程的框架。

注 1：包括需求分析、设计与开发、验证与确认、部署、运维与监测、再评价直至停运。

注 2：在人工智能医疗器械生存周期中，某些活动可出现在不同的过程中，个别过程可重复出现。例如，为了修复系统的隐错和更新系统，需要反复实施开发过程和部署过程。

3.2　数据集术语

3.2.1

数据　data

信息的可再解释的形式化表示，以适用于通信、解释或处理。

注：可以通过人工或自动手段处理数据。

[来源：GB/T 5271.1—2000，01-01-02]

3.2.2

个人敏感数据　personal sensitive data

一旦泄露、非法提供或滥用可能危害人身和财产安全，极易导致个人名誉、身心健康受到损害或歧视性待遇等的个人信息。

注：个人敏感信息包括身份证件号码、个人生物识别信息、银行账号、通信记录和内容、财产信息、征信信息、行踪轨迹、住宿信息、健康生理信息、交易信息、14 岁以下(含)儿童的个人信息等。

[来源：GB/T 35273—2020，3.2，有修改]

3.2.3

健康数据　health data

与身体或心理健康相关的个人敏感数据。

注：由于目前全球规定了不同的隐私合规性法律和法规。例如，在欧洲，可能需要采取的要求和参考变更为“个人数据”和“敏感数据”，在美国，健康数据可能会变更为“受保护的健康信息(PHI)”，这需要不同国家或地区的制造商进一步考虑中国当地的法律或法规。

[来源：IEC/TR 80001-2-2：2012，3.7，有修改]

3.2.4

缺失数据　missing data

按照研究方案要求收集但未观测到的数据。

3.2.5

数据元　data element

用一组属性规定其定义、标识、表示和允许值等的数据单元。

[来源：WS/T 305—2009，3.1.6，有修改]

3.2.6

元数据　metadata

定义和描述其他数据的数据。

[来源：GB/T 18391.1—2009，3.2.16]

3.2.7

数据质量　data quality

在指定条件下使用时，数据的特性满足明确的、隐含的要求的程度。

[来源：GB/T 25000.12—2017，4.3，有修改]

3.2.8

数据集　data set

具有一定主题，可以标识并可以被计算机化处理的数据集合。

[来源：WS/T 305—2009，3.1.2]

3.2.9

训练集　training set

用于训练人工智能算法的数据集，其外部知识源可用于算法参数的计算。

3.2.10

调优集　tuning set

用于优化人工智能算法的数据集，其外部知识源可用于算法超参数的选择。

注：为避免与医疗器械领域所用术语“确认”进行区分，这里不使用通用人工智能领域的 validation set，二者含义一致。

3.2.11

测试集　testing set

用于测试人工智能算法性能的数据集，其外部知识源可用于对算法的评估。

3.2.12

数据生存周期　data life cycle

数据获取、存储、整合、分析、应用、呈现、归档和销毁等各种生存形态演变的过程。

[来源：GB/T 34960.5—2018，3.7]

3.2.13

数据集偏倚　dataset bias

数据集偏离预设目标的一种系统偏差。

3.2.14

数据质量特性　data quality characteristic

对数据质量有影响的数据质量属性的类别。

[来源：GB/T 25000.12—2017，4.4]

3.2.15

数据层次　data stratification

从不同粗细粒度表征的数据层次结构。

注：一般包含以下层级：

——数据元素；

——记录（数据元素的汇集）；

——数据集（记录的汇集）；

——多数据集（数据集的汇集）。

[来源：GB/T 35295—2017，2.1.51，有修改]

3.2.16

参考标准　reference standard

筛查、诊断和治疗过程或基于标注过程建立的基准。

注：参考标准可包含疾病、生理状态或生理异常以及位置和程度等信息标签。

3.2.17

金标准　gold standard

筛查、诊断和治疗可依据的最佳参考标准。

3.2.18

GT 值　ground truth;GT

用于和算法结果进行比对的外部知识源。

3.2.19

数据清洗　data cleaning

检测和修正数据集合中错误数据项的预处理过程。

3.2.20

数据治理　data governance

数据资源及其应用过程中相关管控活动、绩效和风险管理的集合。

[来源:GB/T 34960.5—2018,3.1]

3.2.21

数据挖掘　data mining

对于定量数据,通过从不同视角和维度分析、分类并总结潜在的联系和影响,以此提取模式的计算过程。

[来源:ISO 16439:2014,3.13]

3.2.22

数据标签　data label

附加到一组数据元素的标识符。

[来源:ISO/IEC 2382:2015,2121626,有修改]

3.2.23

数据采集　data acquisition

数据由生成装置按照数据采集规范生成,以数字化格式存储并传输到目标系统的过程。

3.2.24

数据脱敏　data masking

通过去标识化或匿名化,实现对个人敏感信息的可靠保护。

3.2.25

匿名化　anonymization

通过对个人信息的技术处理,使得个人信息主体无法被识别或者关联,且处理后的信息不能被复原的过程。

[来源:GB/T 35273—2020,3.14]

3.2.26

去标识化　de-identification

通过对个人信息的技术处理,使其在不借助额外信息的情况下,无法识别或者关联个人信息主体的过程。

注:去标识化建立在个体基础上,保留了个体颗粒度,采用假名、加密、哈希函数等技术手段替代对个人信息的标识。

[来源:GB/T 35273—2020,3.15]

3.2.27

数据标注　data annotation

对数据进行分析，添加外部知识的过程。

3.2.28

仲裁　arbitration

多名标注人员对同一原始数据的标注结果不一致时用于决定最终结果的过程。

3.2.29

图像　image

物体通过某种成像原理而重显的影像，并可数字化。

注：图像一般以多维坐标系中像素点的位置对应实体点在空间中的位置，以像素点的像素值（或颜色）对应实体点的性质。如放射影像、分子影像、热影像、显微图像等。

[来源：GB/T 34952—2017，2.4，有修改]

3.2.30

图形　graphics

用来表示一个变量相对于其他变量变化情况的线条图，或用以代替文字说明一个概念和思想的图解和表格。

注：如生理信号、基因组图谱等。

[来源：GB/T 34952—2017，2.3]

3.2.31

文本　text

以字符、符号、字、短语、段落、句子、表格或者其他字符排列形式出现的数据，旨在表达一个意义，其解释主要以读者对某种自然语言或人工语言的了解为基础。

注：如电子病历的文本内容等。

[来源：GB/T 5271.1—2000，01.01.03]

3.2.32

数值　numerical value

量化的数值或分类，用来表示水平或状态。

注：如生理状态、生化指标等。

3.2.33

音频　audio

一种数字化动态媒体形态，用于描述声音及其时序性质，并可进行数字人工合成。

[来源：GB/T 34952—2017，2.5，有修改]

3.2.34

视频　video

一种数字化动态媒体形态，用于描述运动图像，进行高速信息传送或显示瞬间的相互关系。

[来源：GB/T 34952—2017，2.6]

注：如内窥镜影像、超声影像等。

3.2.35

多媒体　multimedia

综合表现文本、图形、图像、音频和视频的信息组合。

[来源:GB/T 34952—2017,2.1]

3.3 质量特性术语

3.3.1

软件质量 software quality

在规定条件下使用时,软件产品满足明确或隐含要求的能力。

[来源:GB/T 25000.1—2021,3.42]

3.3.2

软件质量保证 software quality assurance

a) 为使某项目或产品遵循已建立的技术需求提供足够的置信度,而必须采取的有计划的和有系统的全部动作的模式。

b) 设计以估算产品开发或制造过程的一组活动。

[来源:GB/T 11457—2006,2.1294,有修改]

3.3.3

性能 performance

系统或部件在给定的约束条件下实现指定功能的程度。

[来源:GB/T 11457—2006,2.1131,有修改]

3.3.4

性能评价 performance evaluation

为确定运行目标达到了何种有效程度而对系统或系统部件的技术评价。

[来源:GB/T 11457—2006,2.1132]

3.3.5

可靠性 reliability

在规定时间间隔内和规定条件下,系统或部件执行所要求功能的能力。

[来源:GB/T 11457—2006,2.1334]

3.3.6

完整性 integrity

保护数据准确性和完备性的性质。

[来源:GB/T 25000.12—2017,4.12,有修改]

3.3.7

一致性 consistency

a) 在文档或系统或系统部件的各部分之间,一致、标准化、无矛盾的程度。

b) 在数据集的各阶段、部分之间,一致、标准化、无矛盾的程度。

[来源:GB/T 11457—2006,2.320,有修改]

3.3.8

重复性 repeatability

由同一操作员按相同的方法、使用相同的测试或测量设施、在短时间间隔内对同一测试/测量对象进行测试/测量,所获得的独立测试/测量结果间的一致程度。

[来源:GB/T 3358.2—2009,3.3.5,有修改]

3.3.9

再现性　reproducibility

由不同的操作员按相同的方法，使用不同的测试或测量设施，对同一测试/测量对象进行观测以获得独立测试/测量结果，所获得的独立测试/测量结果间的一致程度。

[来源：GB/T 3358.2—2009，3.3.10，有修改]

3.3.10

可达性　accessibility

组成软件的各部分便于选择使用或维护的程度。

[来源：GB/T 11457—2006，2.20]

3.3.11

可得性　availability

a)　软件（系统或部件）在投入使用时可操作或可访问的程度或能实现其指定系统功能的概率。

b)　系统正常工作时间和总的运行时间之比。

c)　在运行时，某一配置项实现指定功能的能力。

[来源：GB/T 11457—2006，2.115，有修改]

3.3.12

保密性　confidentiality

数据对未授权的个人、实体或过程不可用或不泄露的特性。

[来源：GB/T 29246—2017，2.12，有修改]

3.3.13

网络安全　cybersecurity

通过采取必要措施，防范对数据、模型等攻击、侵入、干扰、破坏和非法使用以及意外事故，使设备处于稳定可靠运行的状态，以及保障数据、模型等的完整性、保密性、可得性的能力。

[来源：GB/T 22239—2019，3.1，有修改]

3.3.14

安全性　safety

免除于不可接受的风险。

[来源：YY/T 0316—2016，2.20]

3.3.15

健壮性　robustness

鲁棒性/稳健性

在存在无效输入或急迫的环境条件下，系统或部件其功能正确的程度。

[来源：GB/T 11457—2006，2.1397，有修改]

3.3.16

泛化能力　generalizability

机器学习算法对陌生样本的适应能力。

3.3.17

响应时间　response time

在给定的测试环境下，对给定数量样本进行运算并获得结果所需要的平均时间。

3.3.18

可追溯性　traceability

系统对其决策过程及输出进行记录的特性。

3.3.19

公平性　fairness

系统做出不涉及喜好和偏袒决策的性质。

3.3.20

可解释性　explainability

以人能理解的方式，对系统决策因素进行说明的能力。

3.4　质量评价术语

3.4.1　评价方式

3.4.1.1

性能测试　performance testing

评价系统或部件与规定的性能需求的依从性的测试行为。

[来源：GB/T 11457—2006，2.1135]

3.4.1.2

独立性能测试　standalone performance testing

通过直接比对模型在没有外界干预的情况下产生的结果和参考标准的结果，评估人工智能医疗器械的性能。

3.4.1.3

判读者性能测试　reader performance testing

通过比对判读者在独立工作和结合模型工作两种状态下判读数据的结果，评估人工智能医疗器械的性能。

3.4.1.4

多判读者多病例研究　multi-reader multi-case study

通过判读人员和病例的某种交叉组合方式开展的判读者性能研究，评估人工智能医疗器械的性能。

3.4.1.5

黑盒测试　black-box testing

忽略系统或部件的内部机制只集中于响应所选择的输入和执行条件产生的输出的一种测试。

[来源：GB/T 11457—2006，2.142、2.669]

3.4.1.6

白盒测试　glass-box testing

侧重于系统或部件内部机制的测试。类型包括分支测试、路径测试、语句测试等。

[来源：GB/T 11457—2006，2.678、2.1604]

3.4.1.7

对抗[措施]　countermeasure

为减小脆弱性而采用的行动、装置、过程、技术或其他措施。

[来源：GB/T 25069—2010，2.1.4]

3.4.1.8

对抗样本　adversarial sample

基于原始数据上添加扰动达到混淆系统判别目的的新样本。

3.4.1.9

对抗测试　adversarial test

使用对抗性样本开展的测试，或采用不同目标样本分布的特选数据作为压力数据集进行的测试。

3.4.2　**评价指标**

3.4.2.1

阳性样本　positive sample

由参考标准确定为带有某一种或几种特定特征的样本。

3.4.2.2

阴性样本　negative sample

除阳性样本以外的样本。

3.4.2.3

真阳性　true positive；TP

被算法判为阳性的阳性样本。

3.4.2.4

假阳性　false positive；FP

被算法判为阳性的阴性样本。

3.4.2.5

真阴性　true negative；TN

被算法判为阴性的阴性样本。

3.4.2.6

假阴性　false negative；FN

被算法判为阴性的阳性样本。

3.4.2.7

目标区域　target region

根据参考标准从原始数据中划分出的若干个包含特定类别目标的最小数据子集（子集元素为像素、体素等）。

3.4.2.8

分割区域　segmentation region

从原始数据中划分出的若干个包含特定类别目标的最小数据子集（子集元素为像素、体素等）。

3.4.2.9

病变定位　lesion localization

算法检出病变位置正确标识出参考标准确定的病变位置。

3.4.2.10

非病变定位　non-lesion localization

算法检出病变位置未能正确标识出参考标准确定的病变所在位置。

3.4.2.11

病变定位率　lesion localization rate

病变定位数量占由参考标准确定的全体病变数量的比例。

3.4.2.12

非病变定位率　non-lesion localization rate

非病变定位数量占全体病例数量的比例，非病变定位率可以大于1。

3.4.2.13

假阳性率　false positive rate

假阳性病例数量（阴性病例中包含非病变定位）占全部阴性病例数量的比例。

3.4.2.14

灵敏度　sensitivity

召回率（查全率）　recall

真阳性样本占全体阳性样本的比例。

3.4.2.15

特异度　specificity

真阴性样本占全体阴性样本的比例。

3.4.2.16

漏检率　miss rate

1减去灵敏度。

3.4.2.17

精确度（查准率）　precision

阳性预测值　positive prediction value

真阳性样本占被算法判为阳性样本的比例。

3.4.2.18

阴性预测值　negative prediction value

真阴性样本占被算法判为阴性样本的比例。

3.4.2.19

准确率　accuracy

算法判断正确的样本占全体样本的比例。

3.4.2.20

F_1 度量　F_1-measure

召回率和精确度的调和平均数。

3.4.2.21

约登指数　Youden index

灵敏度与特异度之和减去1。

3.4.2.22

受试者操作特征曲线　receiver operating characteristics curve；ROC curve

以假阳性率为横坐标、真阳性率为纵坐标，根据算法在不同阈值设定下对于给定的测试集得到的一系列结果绘制的曲线。

3.4.2.23

自由响应受试者操作特征曲线　free-response receiver operating characteristics curve；FROC curve

以非病变定位率为横坐标、病变定位率为纵坐标，根据算法在不同阈值设定下对于给定的测试集得到的一系列结果绘制的曲线。

3.4.2.24

候选自由受试者操作特征曲线　alternative free receiver operating characteristics curve；AFROC curve

以假阳性率为横坐标、病变定位率为纵坐标，根据算法在不同阈值设定下对于给定的测试集得到的一系列结果绘制的曲线。

3.4.2.25

精确度-召回率曲线　precision-recall curve；P-R curve

以召回率为横坐标、精确度为纵坐标，根据算法在不同阈值设定下对于给定的测试集得到的一系列结果绘制的曲线。

3.4.2.26

曲线下面积　area under curve；AUC

曲线下与坐标轴围成的积分面积。

3.4.2.27

平均精确度　average precision；AP

精确度-召回率曲线下与坐标轴围成的积分面积。

3.4.2.28

平均精确度均值　mean average precision；MAP

在多目标检测问题上，算法对于各类目标的平均精确度的平均值。

3.4.2.29

交并比　intersection over union；IoU

分割区域与目标区域的交集占分割区域与目标区域并集的比例。

注：也可称为 Jaccard 系数。

3.4.2.30

Dice 系数　Dice coefficient

分割区域与目标区域的交集占分割区域与目标区域平均值的比例。

3.4.2.31

中心点距离　central distance

分割区域中心与目标区域中心的距离，该指标反映两个集合的接近程度。

3.4.2.32

混淆矩阵　confusion matrix

一种矩阵，它按一组规则记录试探性实例的正确分类和不正确分类的个数。

注 1：通常矩阵的列代表人工智能的分类结果，而矩阵的行代表参考标准的分类结果，附录 A 给出示例。

注 2：也可称为含混矩阵。

[来源：GB/T 5271.31—2006，31.02.18，有修改]

3.4.2.33

Kappa 系数　Kappa coefficient

一种用于评价结果一致性的指标。

3.4.2.34

信噪比　signal-to-noise ratio;SNR

信号平均功率水平与噪声平均功率水平的比值。

3.4.2.35

峰值信噪比　peak signal-to-noise ratio

信号最大可能功率水平与噪声平均功率水平的比值。

3.4.2.36

结构相似性　structural similarity

一种衡量两幅图像相似度的指标。

3.4.2.37

余弦相似度　cosine similarity

通过测量两个向量的夹角的余弦值来度量它们之间的相似性。

3.4.2.38

困惑度　perplexity

度量概率分布或概率模型的预测结果与样本的契合程度,困惑度越低则契合越准确。

3.4.2.39

字错率　word error rate

将识别出来的字需要进行修改的字数与总字数的比值。

3.4.2.40

交叉熵　cross-entropy

一种度量两个概率分布之间差异的指标。

3.4.2.41

互信息　mutual information

对两个随机变量间相互依赖性的量度。

3.4.2.42

服务可用性　service availability

服务客户发起服务请求后,服务可访问的时间占总服务时间的比例。

注:服务可用性的计算是在一系列预定义的时间段中,服务可用时间之和占预定义时间段之和的比例,可排除允许的服务不可用时间。

3.5　应用场景术语

3.5.1

计算机辅助　computer-aided

涉及使用计算机完成部分工作的技术或过程。

[来源:ISO/IEC 2382:2015,2121395]

3.5.2

计算机辅助诊断　computer-aided diagnosis

辅助判断患者是否患病、疾病的类型、严重程度、发展阶段、干预措施等。

注:计算机辅助诊断旨在提供除计算机辅助检测结果之上的额外信息,这些信息包含对患者是否患病、疾病的类型、严重程度、发展阶段、干预措施等作出的判断。

3.5.3

计算机辅助检测　computer-aided detection

通过检测、标记、强调或其他方式辅助医务人员注意医疗数据的可能异常情况。其结果供医务人员参考。

3.5.4

计算机辅助分诊　computer-aided triage

自动分析医疗数据、给出初始解释和鉴别分类、辅助医务人员确定患者优先级。

注：导诊不属于计算机辅助分诊。

3.5.5

临床决策支持　clinical decision support

根据临床知识和患者数据产生辅助决策的建议，该建议由医务人员使用。

注：在不同的国家和地区，临床决策支持系统可能不属于医疗器械。

3.5.6

患者决策辅助　patient decision assistant

向患者提供建议或辅助决策，该决策由非医务人员使用，结果仅供参考。

注：在不同的国家和地区，患者决策辅助系统可能不属于医疗器械。

3.5.7

计算机视觉　computer vision

人工视觉　artificial vision

功能单元获取、处理和解释可视数据的能力。

[来源：GB/T 5271.28—2001，28.01.19]

3.5.8

语音识别　speech recognition

自动语音识别　automatic speech recognition；ASR

通过功能单元对人的语音所表示信息的感知与分析。

[来源：GB/T 5271.28—2001，28.01.15]

3.5.9

自然语言处理　natural language processing

自然语言理解和生成及其衍生技术，以从文本化的人类语言中获取有意义的信息。

3.5.10

知识图谱　knowledge graph

将海量知识及其相互联系组织在一张大图中，用于知识的管理、搜索和服务。

3.5.11

医学图像处理　medical image processing

一类对医学图像进行处理的方法。

注：包括图像重建、成像加速、图像增强、图像恢复（降噪、去伪影）、图像分割、图像配准、图像识别、图像分类、目标检测、图像映射、图像可视化等。根据医疗器械应用场景可分为前处理应用和后处理应用。

附 录 A
（资料性）
评价指标计算公式说明

A.1 辅助诊断性能

A.1.1 混淆矩阵

人工智能医疗器械辅助诊断功能多涉及分类问题，算法将患者数据输出两个或两个以上互斥（不相交）的类别或状态。根据输出的类别或状态的数量构成混淆矩阵，通过混淆矩阵的各个参量来评价辅助诊断性能。

注 1：如果算法输出两个或两个以上不互斥（相交）的类别或状态时，可拆分成多个二分类类别构成混淆矩阵（见表 A.2，阴性与阳性根据产品功能特性定义）进行评价。

注 2：对于多个二分类混淆矩阵，例如进行多次测试，或是在多个数据集上进行测试，甚或是执行多分类任务，每两两类别的组合都对应一个混淆矩阵，在 n 个二分类混淆矩阵上综合考虑各项指标时，一种直接的做法是先在各混淆矩阵上分别计算，再计算平均值；还可先将各类混淆矩阵的对应元素进行平均，得到 TP、FP、TN、FN 的平均值再进行计算。

对于分类问题，混淆矩阵的一般形式如表 A.1 所示。

表 A.1 n 分类混淆矩阵

分类	Pred_1	Pred_2	…	…	…	Pred_n
True_1	$N_{1,1}$	$N_{1,2}$	…	…	…	…
True_2	…	$N_{2,2}$	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
True_n	…	…	…	…	…	$N_{n,n}$

注：Pred_x（$x=1\sim n$）为人工智能分类为 x 类的类别；True _x（$x=1\sim n$）为参考标准分类为 x 类的类别；$N_{i,j}$（$i=1\sim n$，$j=1\sim n$）为参考标准的分类结果为 i 类，被人工智能分类为 j 类的个数；n 为分类类型个数。

二分类的混淆矩阵可简化为表 A.2 所示。

表 A.2 二分类混淆矩阵

分类		人工智能分类	
		阳性	阴性
参考标准分类	阳性	TP	FN
	阴性	FP	TN

多分类实际可转化为二分类问题，参考标准分类为 i 类与其他类别的混淆矩阵简化形式如表 A.3 所示。

表 A.3　多分类实际可转化为二分类混淆矩阵

分类		人工智能分类	
		阳性	阴性
参考标准分类	阳性	$TP=\sum_{i=1}^{n}N_{i,i}$	$FN=\sum_{j=1,j\neq i}^{n}N_{i,j}$
	阴性	$FP=\sum_{j=1,j\neq i}^{n}N_{j,i}$	$TN=\sum_{j=1,j\neq i}^{n}\sum_{l=1,l\neq i}^{n}N_{j,l}$

A.1.2　灵敏度

灵敏度用 Sen 表示，表达式见公式(A.1)：

$$Sen=\frac{TP}{TP+FN}\times 100\% \quad \cdots\cdots (A.1)$$

式中：

Sen ——灵敏度；

TP ——真阳性样本的个数；

FN ——假阴性样本的个数。

A.1.3　特异度

特异度用 Spe 表示，表达式见公式(A.2)：

$$Spe=\frac{TN}{FP+TN}\times 100\% \quad \cdots\cdots (A.2)$$

式中：

Spe ——特异度；

TN ——真阴性样本的个数；

FP ——假阳性样本的个数。

A.1.4　漏检率

漏检率用 MR 表示，表达式见公式(A.3)：

$$MR=1-Sen \quad \cdots\cdots (A.3)$$

式中：

Sen ——灵敏度；

MR ——漏检率。

A.1.5　阳性预测值

阳性预测值用 PPV 表示，表达式见公式(A.4)：

$$PPV=\frac{TP}{TP+FP} \quad \cdots\cdots (A.4)$$

式中：

PPV ——阳性预测值；

TP ——真阳性样本的个数；

FP ——假阳性样本的个数。

A.1.6 阴性预测值

阴性预测值用 NPV 表示，表达式见公式(A.5)：

$$NPV = \frac{TN}{FN + TN} \quad \cdots\cdots (A.5)$$

式中：

NPV ——阴性预测值；

TN ——真阴性样本的个数；

FN ——假阴性样本的个数。

A.1.7 准确率

准确率用 Acc 表示，表达式见公式(A.6)：

$$Acc = \frac{\sum_{i=1}^{n} N_{i,i}}{\sum_{j=1}^{n}\sum_{l=1}^{n} N_{j,l}} \quad \cdots\cdots (A.6)$$

A.1.8 约登指数

约登指数用 Y 表示，表达式见公式(A.7)：

$$Y = Sen + Spe - 1 \quad \cdots\cdots (A.7)$$

A.1.9 Kappa 系数

Kappa 系数用 K 表示，表达式见公式(A.8)：

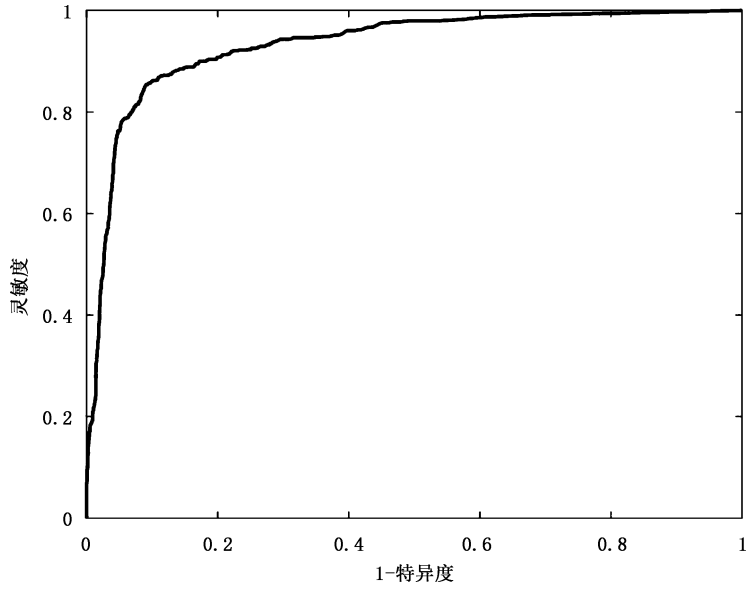
$$K = \frac{Acc - p_e}{1 - p_e} \quad \cdots\cdots (A.8)$$

式中 p_e 的表达式见公式(A.9)：

$$p_e = \frac{\sum_{i=1}^{n}(\sum_{j=1}^{n} N_{i,j} \times \sum_{j=1}^{n} N_{j,i})}{(\sum_{a=1}^{n}\sum_{b=1}^{n} N_{a,b})^2} \quad \cdots\cdots (A.9)$$

A.1.10 受试者操作特征曲线

受试者操作特征曲线如图 A.1 所示。



横坐标和纵坐标值均为0～1之间，曲线下面积AUC最大值为1。

图 A.1 ROC曲线

A.2 辅助检测性能

A.2.1 病变定位率

人工智能医疗器械的辅助检测功能指的是标出数据中的目标作为给医生的提示，如在医学影像中找到病灶。辅助检测的结果举例如下：在 N 张图像中，共含有Les个病灶，其中LL个病灶被人工智能算法正确检出，NLL个不含有病灶的图像区域被误诊，N 个病例中有 M 个病例包含非病变定位。

病变定位率用LLR表示，表达式见公式(A.10)：

$$LLR = \frac{LL}{Les} \times 100\% \quad \cdots\cdots (A.10)$$

式中：

LLR ——病变定位率；

LL ——算法检出病变位置正确标识出参考标准确定的病变位置的数量；

Les ——为参考标准确定的全体病变数量。

注：病变定位率也可称为召回率，定义见3.4.2.14，LL对应真阳性样本，Les对应全体阳性样本。

A.2.2 精确度

精确度用Pre表示，表达式见公式(A.11)：

$$Pre = \frac{LL}{LL + NLL} \times 100\% \quad \cdots\cdots (A.11)$$

式中：

Pre ——精确度；

LL ——算法检出病变位置正确标识出参考标准确定的病变位置的数量；

NLL ——算法检出病变位置未能正确标识出参考标准确定的病变位置的数量。

注：精确度的定义见 3.4.2.17，LL 对应真阳性样本，LL 与 NLL 之和对应被算法判为阳性的样本。

A.2.3 非病变定位率

非病变定位率用 NLR 表示，表达式见公式(A.12)：

$$NLR=\frac{NLL}{N}\times 100\% \quad \cdots\cdots(A.12)$$

式中：

NLR ——非病变定位率；

NLL ——算法检出病变位置未能正确标识出参考标准确定的病变位置的数量；

N ——全体病例数量。

注：也可称为单病例平均假阳个数。

A.2.4 假阳性率

假阳性率用 FPR 表示，表达式见公式(A.13)：

$$FPR=\frac{M}{N_N}\times 100\% \quad \cdots\cdots(A.13)$$

式中：

FPR ——假阳性率；

M ——被判为假阳性病例的阴性病例总数。在计算候选自由受试者操作特征曲线时，每个阴性病例如含有非病变定位且其最大值高于决策阈值，视为假阳性病例，反之视为阴性。

N_N ——阴性病例总数。

A.2.5 F_1 度量

F_1 度量用 F_1 表示，表达式见公式(A.14)：

$$F_1=\frac{2\times LLR\times Pre}{LLR+Pre} \quad \cdots\cdots(A.14)$$

式中：

F_1 ——F_1 度量；

LLR ——病变定位率；

Pre ——精确度。

注：F_1-measure 的一般形式 F_β-measure 表示召回率和精确度的不同权重，定义为加权调和平均数，用公式(A.15)表示。

$$F_\beta=\frac{(1+\beta^2)\times Pre\times LLR}{\beta^2\times Pre+LLR} \quad \cdots\cdots(A.15)$$

式中：

F_β ——F_β 度量；

LLR ——病变定位率；

Pre ——精确度；

β ——权重参数，当 $\beta>1$ 时，召回率的权重大于精确度；$\beta=1$ 时，召回率与精确度的权重相等；$\beta<1$ 时，精确度的权重大于召回率。

A.2.6 自由响应受试者操作特征曲线

自由响应受试者操作特征曲线如图 A.2 所示。

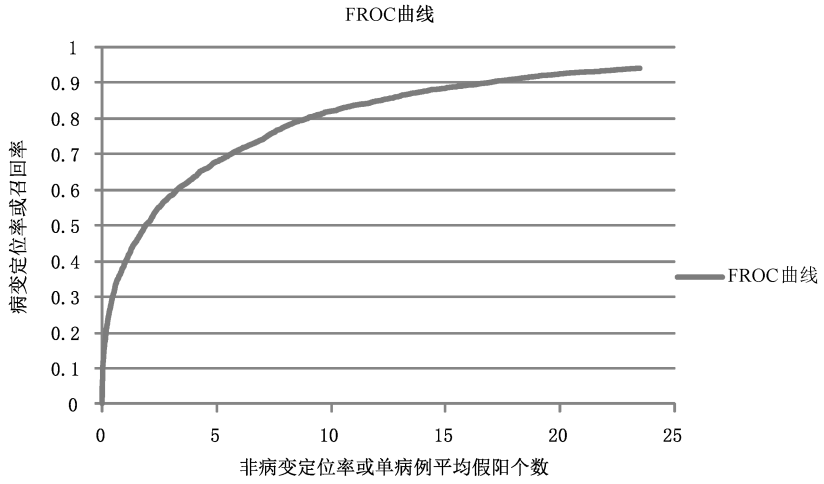


图 A.2 FROC 曲线

A.2.7 精确度-召回率曲线

精确度-召回率曲线如图 A.3 所示。

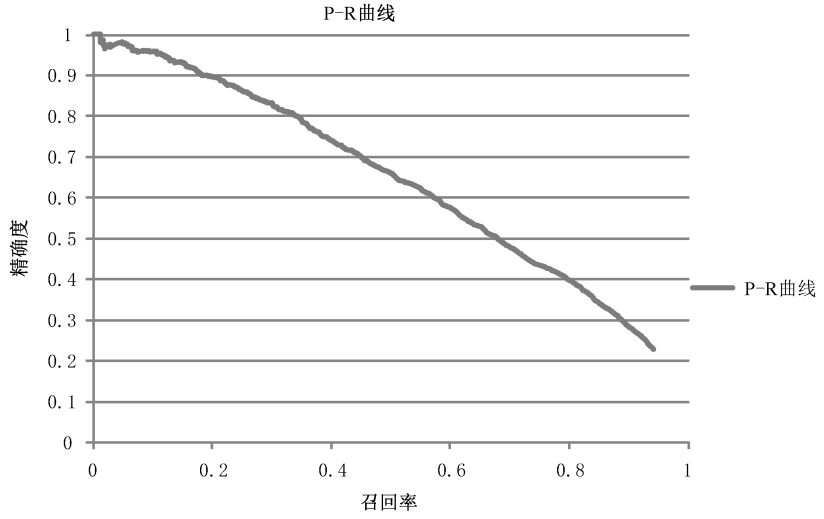


图 A.3 P-R 曲线

A.2.8 平均精确度

平均精确度用 AP 表示，表达式见公式(A.16)：

$$AP=\int_0^1 Pre(Rec)d(Rec) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (A.16)$$

式中：

AP ——平均精确度；

Pre ——精确度；

Rec ——召回率。

A.2.9 平均精确度均值

平均精确度均值用 MAP 表示，表达式见公式(A.17)：

$$MAP=\frac{\sum_{n=1}^{N}AP(n)}{N} \quad \cdots\cdots(A.17)$$

式中：

MAP ——平均精确度均值；

AP ——平均精确度；

N ——类别的数量；

n ——类别 n。

A.3 辅助分割性能

A.3.1 召回率

人工智能医疗器械辅助分割功能多为对医学影像病变进行识别检出，并进行病变、器官或组织边界的分割，如图 A.4 所示，A 为目标区域，B 为分割区域，C 为二者的交集($C=A\cap B$)。

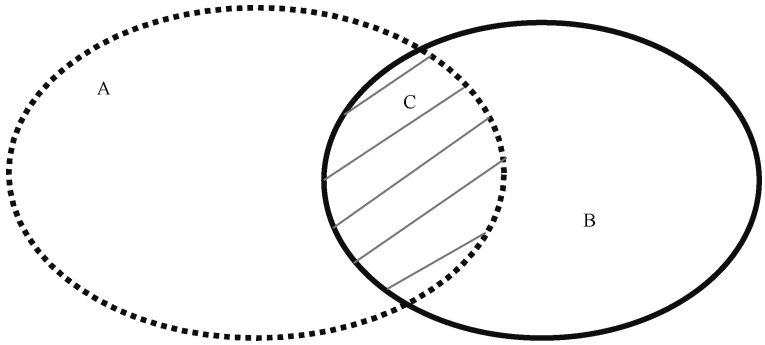


图 A.4 辅助分割示意图

召回率用 Rec 表示，表达式见公式(A.18)：

$$Rec=\frac{|A\cap B|}{|A|} \quad \cdots\cdots(A.18)$$

式中：

Rec ——召回率；

A ——目标区域；

B ——分割区域。

注：召回率的定义见 3.4.2.14，分割区域与目标区域的交集对应真阳性样本，目标区域对应全体阳性样本。

A.3.2 精确度

精确度用 Pre 表示，表达式见公式(A.19)：

$$Pre = \frac{|A \cap B|}{|B|} \quad \cdots\cdots (A.19)$$

式中：

Pre ——精确度；

A ——目标区域；

B ——分割区域。

注：精确度的定义见 3.4.2.17，分割区域与目标区域的交集对应真阳性样本，分割区域对应被算法判为阳性的样本。

A.3.3 Dice 系数

Dice 系数用 Dice 表示，表达式见公式(A.20)：

$$Dice = 2 \times \frac{|A \cap B|}{|A| + |B|} \quad \cdots\cdots (A.20)$$

式中：

Dice ——Dice 系数；

A ——目标区域；

B ——分割区域。

A.3.4 Jaccard 系数

Jaccard 系数用 Jaccard 表示，表达式见公式(A.21)：

$$Jaccard = \frac{|A \cap B|}{|A \cup B|} \quad \cdots\cdots (A.21)$$

式中：

Jaccard ——Jaccard 系数；

A ——目标区域；

B ——分割区域。

参 考 文 献

[1] GB/T 3358.2—2009 统计学词汇及符号 第2部分:应用统计

[2] GB/T 5271.1—2000 信息技术 词汇 第1部分:基本术语

[3] GB/T 5271.28—2001 信息技术 词汇 第28部分:人工智能 基本概念与专家系统

[4] GB/T 5271.31—2006 信息技术 词汇 第31部分:人工智能 机器学习

[5] GB/T 5271.34—2006 信息技术 词汇 第34部分:人工智能 神经网络

[6] GB/T 11457—2006 信息技术 软件工程术语

[7] GB/T 18391.1—2009 信息技术 元数据注册系统(MDR) 第1部分:框架

[8] GB/T 22239—2019 信息安全技术 网络安全等级保护基本要求

[9] GB/T 25000.1—2021 系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第1部分:SQuaRE指南

[10] GB/T 25000.12—2017 系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第12部分:数据质量模型

[11] GB/T 25069—2010 信息安全技术 术语

[12] GB/T 29246—2017 信息技术 安全技术信息安全管理体系 概述和词汇

[13] GB/T 32400—2015 信息技术 云计算 概览与词汇

[14] GB/T 34952—2017 多媒体数据语义描述要求

[15] GB/T 34960.5—2018 信息技术服务 治理 第5部分:数据治理规范

[16] GB/T 35273—2020 信息安全技术 个人信息安全规范

[17] GB/T 35295—2017 信息技术 大数据 术语

[18] WS/T 305—2009 卫生信息数据集元数据规范

[19] YY/T 0316—2016 医疗器械 风险管理对医疗器械的应用

[20] YY/T 0664—2020 医疗器械软件 软件生存周期过程

[21] ISO/IEC 2382:2015 Information technology—Vocabulary

[22] ISO 16439:2014 Information and documentation—Methods and procedures for assessing the impact of libraries

[23] IEC/TR 80001-2-2:2012 Application of risk management for IT-networks incorporating medical devices—Part 2-2: Guidance for the communication of medical device security needs, risks and controls

[24] 周志华.机器学习[M].北京:清华大学出版社,2016

[25] 方积乾.卫生统计学.第7版[M].北京:人民卫生出版社,2012

[26] 全国科学技术名词审定委员会.计算机科学技术名词[M].北京:科学出版社,2018

索　　引

汉语拼音索引

英文对应词索引

A

B

C

D

E

F

G

H

I

K

L

M

N

O

P

R

S

T

U

V

W

Y